

Multimodal Evaluation for Medical Image Segmentation

Rubén Cárdenes¹, Meritxell Bach², Ying Chi⁵, Ioannis Marras⁴, Rodrigo de Luis¹, Mats Anderson³, Peter Cashman⁵ and Matthieu Bultelle⁵

¹ LPI, University of Valladolid, Spain

² EPFL, Lausanne, Switzerland

³ MI, Linköping University, Sweden

⁴ Aristotle University of Thessaloniki, Greece

⁵ Imperial College, London, UK

Abstract. This paper is a joint effort between five institutions that introduces several novel similarity measures and combines them to carry out a multimodal segmentation evaluation. The new similarity measures proposed are based on the location and the intensity values of the misclassified voxels as well as on the connectivity and the boundaries of the segmented data. We show experimentally that the combination of these measures improve the quality of the evaluation. The study that we show here has been carried out using four different segmentation methods from four different labs applied to a MRI simulated dataset of the brain. We claim that our new measures improve the robustness of the evaluation and provides better understanding about the difference between segmentation methods.

Key words: Multimodal evaluation, Segmentation, Similarity Measures, Brain tissue segmentation

1 Introduction and State of the Art

The goal of medical image segmentation is to obtain a labeled image where each label corresponds to the real anatomy of the patient. Several technical factors make this goal hard if not impossible to achieve with the current technology, therefore measurements of the quality of the results are needed to compare segmentation methods.

Many works to evaluate segmentation methods has been reported in the last two decades. A good survey about them can be found in [1]. This author distinguishes the evaluation methods between empirical (based on the study of the results) and analytical (based only on intrinsic features of the methods). The empirical methods are divided into goodness and discrepancy methods, where the former are based on the study of the results themselves, and the latter compare the results with a reference or ground truth. Among the discrepancy methods, there exist several features reported to measure the quality of the segmentation: number of misclassified voxels, position of misclassified voxels, number of objects in the image, feature values of segmented objects and other miscellaneous quantities.

Most of the methods in the literature for segmentation evaluation are based on classic discrepancy methods, limited to the computation of the number of voxels of the segmented classes in the results and in a gold standard. Other authors have introduced the location of the misclassified voxels as a feature to measure the discrepancy between segmented images, for example, Yasnoff [2], Straters [3] and later Pichon [4] proposed to use an error distance from the misclassified voxels to the gold standard. Huttenlocher [5] use the partial Hausdorff distance between set of voxels, and also [6] proposed an overlap distance using fuzzy set theory to take into account fractional labels coming from multiple test images. Other work proposed by Cardoso [7] presents a general distance between segmentation partitions to measure the quality of a given segmentation.

One interesting work about segmentation evaluation is the one published by Udupa [8] who proposed a methodology based on several features, not only on the accuracy of the segmentation, but also on reproducibility and efficiency, and he stated that the combination of those factors are essential in the assessment of the performance of any segmentation method.

The main goal of this paper is to introduce new similarity measures and to show that their combination will improve the quality of segmentation evaluation, in terms of accuracy, using a known ground truth. In order to show this we will compare four segmentation techniques for a specific application: brain tissue segmentation. There is of course, a problem inherent to this way of evaluation, because it is quite difficult to obtain a reliable reference segmentation dataset. The most used approach is to use manual segmentation, or a combination of several manual segmentations, from several experts if possible. There is however the possibility to validate brain tissue segmentation methods on a brain *simulated* data set as the one proposed by the *Brain Web* MR simulator [9]. Their data is very well-suited for this purpose since a ground-truth classification is known.

2 Evaluation Study

Notice that our goal is to show our multimodal evaluation method, not to validate any method used here, so the methods used are not of relevant importance. The methods selected to perform the study are:

- Mean-Shift initialized Level Set method, (**MSiLS**)
- Statistical Parametric Classification using Gaussian Hidden Markov Random Field Model, (**GHMRF**)
- k Nearest Neighbors, (**kNN**)
- Split and Merge Segmentation, (**SM**)

As we have said, in this work we use the images from the Brain Web MR simulator [9], particularly the dataset with noise 5% and no RF on the T1-weighted modality. The volume used has been preprocessed to remove non brain tissues.

We show in Fig. 1, the segmentation results for an axial slice, using blue for CSF, yellow for GM and dark green for WM. Overlapping between GM and WM is shown in grey and pink, and in blue and red the voxels the overlap between CSF and GM. There is no overlap between CSF and WM in the slice shown.

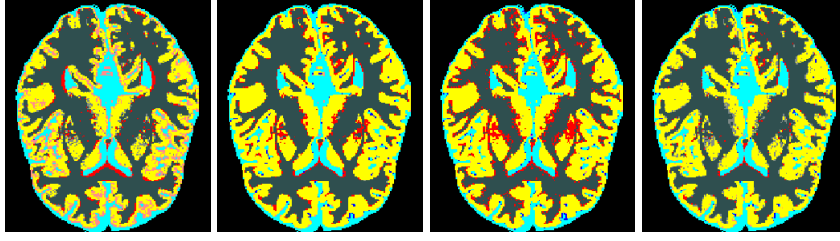


Fig. 1. Segmented axial slices with error voxels overlapped, using from left to right: MSiLS, GHMRF, kNN and SM methods

2.1 Classic Similarity Measures

One classic approach to determine how good are the segmentations, are similarity measures based on region overlap. One of the most common measures is the construction of the confusion tables, whose values represent the overlapping between two classes with respect to the number of voxels of the class in the gold standard. Other common measures used are the Jaccard (JC), Dice Similarity (DS), Tanimoto (TN), and Volume Similarity (VS) coefficients. All of them take values between 0 and 1. If X is the set of voxels segmented as class c in one volume, Y is the set of voxels of the same class in the other volume, $a = |X \cap Y|$, $b = |X \setminus Y|$, $c = |Y \setminus X|$, $d = |\overline{X \cup Y}|$, and $|\cdot|$ stands for the number of elements, these measures are defined with the following expressions,

$$JC := \frac{|X \cap Y|}{|X \cup Y|} = \frac{a}{a + b + c} \quad (1)$$

$$DS := \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2a}{2a + b + c} \quad (2)$$

These two coefficients are equal to one if X and Y are the same region, and zero if they are disjoint regions. In fact, they are related by $DS = 2JC/(JC + 1)$.

$$TN := \frac{|X \cap Y| + |\overline{X \cup Y}|}{|X \cup Y| + |\overline{X \cap Y}|} = \frac{a + d}{a + 2b + 2c + d} \quad (3)$$

This coefficient is one if X is equal to Y , and zero if they are disjoint regions and they occupy all the image.

$$VS := 1 - \frac{||X| - |Y||}{|X| + |Y|} = 1 - \frac{|b - c|}{2a + b + c} \quad (4)$$

This is one if $|X| = |Y|$, and zero if one of the regions is empty. In Fig. 2 we show the results of these similarity measures computed over the segmented volumes obtained with each method.

Looking at Fig. 2, we can have a rough idea about the accuracy of the different methods. However, some measures like the TN coefficients differ from the values

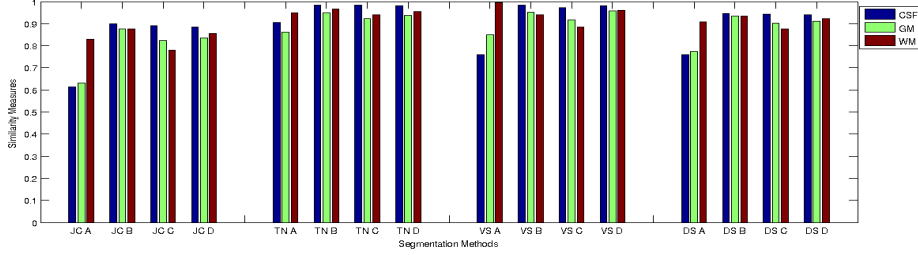


Fig. 2. Classic similarity measures (JC, TN, VS and DS) computed for all methods, A: MSiLS, B: GHMRF, C: kNN, and D: SM

obtained by the other coefficients (the classes are ordered different than in the other three coefficients) and produce values that hardly can differentiate the methods. This is because it depends on the number of voxels outside X and Y , that can be very large in our case, therefore leading to values near one, even if there is not too much overlapping. The VS coefficients present results not realistic (notice an almost perfect classification of WM in MSiLS method), that is because it depends only on the number of voxels of X and Y , and it can be one even if there exists no overlapping at all. Finally the JC and VS coefficients show equivalent values as expected. For those reasons, we will use the JC coefficient for our evaluation study.

2.2 Distance Based Similarity Measures

The similarity measures described above are based only on the size of classified regions. We propose in this section to include the voxels location to improve qualitatively the measures. We can define the distances from the misclassified voxels as in [4]

$$d(r) := \begin{cases} 0, & r \in X \cap Y \\ \min_{x \in X} \|r - x\|, & r \in Y \setminus X \\ \min_{y \in Y} \|r - y\|, & r \in X \setminus Y \end{cases} \quad (5)$$

We propose to use this distance to define a new similarity measure that takes values between 0 and 1. The idea is to penalize more those voxels that are more distant from their corresponding class in the gold standard, i.e. to weight every misclassified voxel by its Euclidean distance to the nearest voxel of the class it should belong to. We will use the squares of the distances to penalize more to very distant voxels.

The new measure we propose is called JCd , and is defined by substituting the values b and c from (1), by $\sum_i d(x_i)^2$ and $\sum_i d(y_i)^2$ respectively, where x_i are misclassified voxels of X that should be classified as Y , y_i are voxels of Y that should be classified as X , and $d()$ is the distance defined in (5). We use the JC coefficient for the reasons commented in sect. 2.1.

2.3 Intensity Based Similarity Measures

In this section we introduce another similarity measure, but using distances in the intensity space. The idea is to penalize more to misclassified voxels that are close to the theoretic mean of a class, because they are supposed to be easy to classify. Therefore, we will define a weighting function, dependent on the theoretic mean and variance of each class, obtained from the gold standard. Defining three Gaussian probability density functions for each class, Y_{csf} , Y_{gm} and Y_{wm} , we define the weighting function F as

$$F = H(1 + Y_{csf} + Y_{gm} + Y_{wm}) \quad (6)$$

where H is a constant that modulates the penalization effect. We show the weighting function F , in Fig. 3 (a).

The new similarity measure, that we will call JCI , is defined changing b and c by $\sum_i F(x_i)$ and $\sum_i F(y_i)$ respectively in (1). Again, we obtain a measure constrained between 0 and 1, and the results obtained are shown in Fig. 3 (b), using $H = 10$.

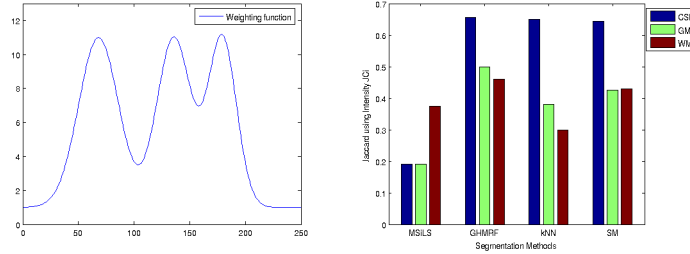


Fig. 3. Weighting function F (left), and intensity based similarity measure computed for all methods and classes (right)

2.4 Connectivity Coefficient

Other similarity measure can be defined using the connectivity of labeled images. We will consider two regions connected if one or more of the 26 neighbors of any voxel in one region belong to the other region. In this case we will compare the number of connected components for each class N_{X_c} in the segmented volume with the number of connected components for the same class in the gold standard N_{Y_c} . The definition of a connectivity coefficient CC that takes values between 0 and 1 can be expressed as

$$CC_c := \frac{\min\{N_{X_c}, N_{Y_c}\}}{N_{X_c} + N_{Y_c}} \quad (7)$$

2.5 Similarity Measures on the Boundaries

It is also interesting to use the segmented boundaries to measure the similarity between the ground truth and the segmentations. A measure between 0 and 1 can be defined using the JC for boundaries. Given the boundary of one segmented class, c , ∂X_c , and the boundary of that class in the ground truth, ∂Y_c , the boundary JC coefficient is defined as

$$BJC_c := \frac{|\partial X_c \cap \partial Y_c|}{|\partial X_c \cup \partial Y_c|} \quad (8)$$

Sometimes, the segmented images may contain many small groups of isolated voxels. Of course, those erroneous voxels are significant on our measures, but we want a measure definition that does not take into account those voxels, because counting scattered voxels will decrease this similarity measure even if the boundary of the ground truth really fits with the boundary of the segmented image. Therefore, we will use a modified boundary for every class in the segmented image $\partial X'(c)$, defined as the boundary voxels except those connected components in $\partial X'(c)$ that does not have any voxel in common with $\partial Y(c)$. Again we use a 26 neighborhood to define connectivity. The modified boundary JC measure is expressed as

$$BJC'_c = \frac{|\partial X'_c \cap \partial Y_c|}{|\partial X'_c \cup \partial Y_c|} \quad (9)$$

2.6 Global Multimodal Similarity Measure

We propose to use the above definitions to combine different features to obtain more objective and reliable results. In this work we state that, as in human vision, an intelligent system should employ several features to decide between different results. An intelligent similarity measure, will emerge from the combination of the measures proposed here: a multimodal similarity measure. Figure 4 illustrates better our idea. In that figure we plotted in 2D, similarity measure values, choosing as the x and y axes, different combinations of similarity measures. Using this representation we can see more clearly the differences between several methods than in unidimensional plots. In the figures we have also plotted circles centered at the middle point of each method, by averaging the values of all classes, and using a radius proportional to the standard deviation. Notice that better measures correspond to smaller circles and closer to the (1, 1) point.

This idea is expressed numerically defining a global similarity measure that includes all the measures described before. Let \mathbf{v}_c be a vector of similarity measures for a class c ,

$$\mathbf{v}_c = [JC_c, JCd_c, JCi_c, CC_c, BJC'_c] \quad (10)$$

We will define the global similarity measure for a given class c , as

$$G_c := \frac{1}{5}[\mathbf{v}_c \mathbf{v}_c^T]^{1/2} \quad (11)$$

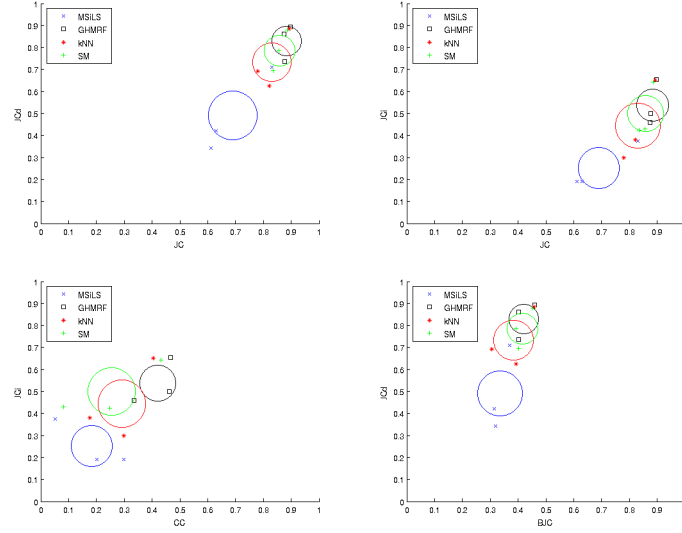


Fig. 4. 2D plots of similarity measures, JC vs JCd (top left), JC vs JCi (top right), CC vs JCi (bottom left) and BJC vs JCd (bottom right)

This measure takes values between 0 and 1, being 0 the worst case and 1 the best case. To obtain a final value for the entire method, we can combine the values obtained for each class, weighting with the number of voxels of each class in the gold standard $|Y_c|$:

$$G := \frac{\sum_c G_c |Y_c|}{\sum_c |Y_c|} \quad (12)$$

We show in Fig. 5, the values for the global similarity measures per class and for the whole segmentation, for the four different methods studied.

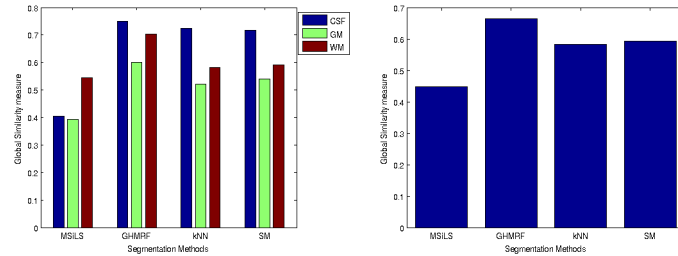


Fig. 5. Global similarity measures per class (left) and averaged (right)

3 Conclusions and Future Works

We have shown that classic similarity measures produce similar values that could arise in erroneous decisions. Therefore, we have proposed a set of new similarity measures and a combination of them, to introduce a new global multimodal similarity measure to obtain better reliability in segmentation evaluation, which is the main contribution of this work. As far as we know, the measures described from section 2.2 are completely new, and this is the first time that multiple similarity measures are combined in this way for segmentation evaluation.

We have also presented 2D plots of pairs of similarity measures that show how the combination of several measures improves the visual representation of the difference between several methods, and motivate the validity of the multi-dimensional or multimodal global measure proposed here.

The correspondence between visual inspection (see Fig. 1), and the numeric similarity measure values fits quite well, presenting good results for GHMRF because it is very well suited for this application, fairly good results for SM and kNN, and SMiLS method performs also good, taking into account that it is not optimized for this task.

The evaluation study done here is not exhaustive, and it should be considered as a good example of how our evaluation method can be applied. Notice also that new measures not related to accuracy, based on reproducibility, efficiency and user interaction, can be included in our model, as proposed by Udupa [8].

Acknowledgments

This work has been funded by the European network of excellence Similar, FP6-507609, to whom all the authors belong to.

References

1. Zhang, Y.: A review of recent evaluation methods for image segmentation. In: Int. Symposium on Signal Proc. and its Applications (ISSPA). (2001) 148–151
2. Yasnoff, W., Miu, J., Bacus, J.: Error measures for scene segmentation. *Pattern Recognition* **9** (1977) 217–231
3. Straters, K., Gerbrands, J.: Three-dimensional segmentation using a split, merge and group approach. *Pattern Recognition Letters* **12** (1991) 307–325
4. Pichon, E., Tannenbaum, A., Kikinis, R.: A statistically based flow for image segmentation. *Medical Image Analysis* **8** (2004) 267–274
5. Huttenlocher, D., Klanderman, G., Rucklidge, W.: Comparing images using the hausdorff distance. *PAMI* **15**(9) (1993) 850–863
6. Crum, W., Camara, O., Hill, D.: Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Tr. Med. Imag.* **25**(11) (2006) 1451–1461
7. Cardoso, J., Corte-Real, L.: Toward a generic evaluation of image segmentation. *IEEE Transactions on Image Processing* **14**(11) (Nov. 2005) 1773–1782
8. Udupa, J., LaBlanc, V., Schmidt, H., Imielinska, C., Saha, P., Grevera, G., Zhuge, Y., Molholt, P., Jin, Y., Currie, L.: A methodology for evaluating image segmentation algorithms. In: SPIE Conf. on Medical Imaging, San Diego, CA, USA (2002)
9. Collins, D., Zijdenbos, A., Kollokian, V., Sled, J., Kabani, N., Holmes, C., Evans, A.: Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging* **17**(3) (1998) 463–468